対話における文脈を考慮した 分散表現生成ニューラルネットワーク†

三上 佳孝*・萩原 将文*

自然言語処理において、文の意味推定は重要な課題の一つである。平叙文中における文の数学的表現としては、共起情報や word2vec に代表される分散表現など様々なものが用いられている。しかし、対話文の特徴を考慮した文の数学的表現はほとんど提案されていない。そこで本論文では、対話文特有の特徴を捉えた分散表現を生成するためのニューラルネットワークを新たに提案する。具体的には、word2vec にて使用されている 2 種類のニューラルネットワークを文生成向けに改良し、組み合わせて使用する。提案ネットワークのうちの一つは再帰型ニューラルネットワークの一種である Long Short-Term Memory をエンコーダとして用い発話文を入力し、そこから前後の発話における Bag-of-Words を推定する学習を行う。もう一つのネットワークは、複数の入力文をエンコーダに入力し次に現れる発話文の Bag-of-Words を推定する学習を行う。これらの学習を行ったエンコーダを組み合わせ、未知の発話文に対する分散表現を生成するニューラルネットワークを構成する。評価実験として、対話破綻検出タスク、ならびに対話行為推定を行い、提案ニューラルネットワークを用いて生成した分散表現の素性としての有効性を検証した。対話破綻検出においては提案手法にて作成した分散表現を素性として用い分類を行うことで、既存手法よりも高い精度が得られた。同様に、対話行為推定においても人手で作成したコーパスを用いて実験を行った。その結果、Bag-of-Words や既存の文ベクトル化手法の一つである Paragraph vector を素性として用いた場合と比較して、提案手法を用いた方がより高い精度を得られた。

キーワード：ニューラルネットワーク、自然言語理解、分散表現、対話行為推定、Long Short-Term Memory

1. はじめに

対話システムや会話型エージェントが近年大きな注目を集めている。ELIZA [1] に代表されるような旧来型の対話システムは対話制御をルールベースで行うものが主流であったが、近年提案されているシステムの多くはユーザーの発話文の解析を行い、その解析結果に対して統計的処理を行っている。数ある統計的手法の中でも特に、ユーザーの発話文をベクトルに変換して処理を行う対話システムは多く提案されている。Banch ら [2] はベクトル空間モデルを素性として用いた非タスク指向型対話システムを提案しており、ルールベースでは行えなかった複雑な対話制御を実現している。また発話文をベクトルに変換して処理を行うシステムの中でも、機械翻訳的手法を応用した対話システムはその柔軟な応答文生成から近年特に注目を集めている。Ritter ら [3] は統計的機械翻訳の手法を応用した対話システムを提案し、従来よりも自然な応答を実現している。また Vinyals ら [4] は機械翻訳のモデルの一種である Sequence-to-Sequence Model [5] を応用した Neural Conversational Model を提案し、より柔軟な応答文生成を実現している。

しかし、Neural Conversational Model をはじめとした統計的手法を用いる対話システムは、対話における文脈を考慮していないものが多く、全く見当外れの返答を返してしまう場合も

少なくない。これは直前のユーザーの発話文のみを入力として用いていることに由来しており、ユーザーの発話とそれに対する応答文の一对一の写像関係のみを学習しているからである。対話における文脈を考慮する方法としては、部分観測マルコフ決定過程（partially-observable Markov decision process：POMDP）を用いる手法 [6, 7] や隠れマルコフモデルを用いる手法 [8]、Recurrent Neural Network（RNN）を用いることで過去の発話を考慮する手法 [9] などが提案されている。しかし POMDP や隠れマルコフモデルを用いる手法は離散的な一意の値として過去の状態を推測し決定する手法である為、ベクトルを入力として内部で連続値に変換するような手法とは相性が悪く、組み合わせて学習を行うのが難しいという弱点が存在する。また RNN についても、学習の複雑さなどから汎用性が低いという弱点が挙げられる。また発話文をベクトルに変換して対話制御を行うモデルにおいて、ベクトルとして文脈を考慮して取得するような手法は提案されていない。

そこで本論文では、対話システムへの入力文を解析する手法として、対話中の文脈を考慮可能な分散表現を生成するニューラルネットワークを提案する。この手法を用いることにより教師なしでより深い前後文脈との関連を考慮したベクトルを取得することが可能となり、ベクトルを入力とする対話システム中などにおいて前後の文脈との関連を考慮して入力文解析を行うことが可能になると考えられる。具体的には、word2vec [10] にて使用されている 2 種類のニューラルネットワークを文生成向けに改良し、これらを組み合わせて使用する。提案ネットワークのうちの一つは再帰型ニューラルネットワークの一種である Long Short-Term Memory（LSTM）をエンコーダとして用い発話文を入力し、そこから前後の発話における Bag-of-Words

† Distributed Representation Generation Neural Network Considering Conversational Context

Yoshitaka MIKAMI and Masafumi HAGIWARA

* 慶應義塾大学大学院理工学研究科
Graduate School of Science and Technology, Keio University

を推定する学習を行う。もう一つのネットワークは、複数の入力文をエンコーダに入力し次に現れる発話文の Bag-of-Words を推定する学習を行う。これらの学習を行ったエンコーダを組み合わせ、未知の発話文に対して分散表現を生成するような新しいニューラルネットワークを構成する。

以下、第2章では関連研究を述べ、第3章で提案するニューラルネットワークを説明する。第4章では評価実験について説明し、第5章で結論を述べる。

2. 関連研究

2.1 対話文の特徴獲得

対話中で発話文が持つ役割を表現する代表的な方法として、対話行為タグを用いる手法が挙げられる。これは発話の話題や種類について分類したもの [11] であり、対話のドメインごとに様々な分類基準が提案されている。Allen ら [12] は発話における表層的な情報を Switchboard Discourse Annotation and Markup System of Labeling (SWBD-DAMSL) タグを用いて表現している。それ以外にも Bracewell ら [13] は社会的側面を重視した対話行為タグを提案しており、Zarischeva ら [14] は Twitter に特化した対話行為タグの分析を行い提案している。

これらの対話行為タグを用いる方法には、複数のタグを用いることによって多様な角度から同一の発話を分析することができるという利点がある。しかし逆に、それぞれのタグは特定の側面からしか特徴を捉えることができず、対話文全体の特徴を捉えることができない。また SWBD-DAMSL タグなどはタグの種類が 48 種類もありコーパスに対してタグ付けを行うコストが高く、結果もアノテータによってばらつきが生まれる等の欠点が存在する。

対話文のみを用いて教師なし学習で文を特徴ごとに分類する手法も提案されている。Ritter ら [15] は Twitter のデータに対して HMM の教師なし学習アルゴリズムを用いて推定を行い、8 つの遷移状態を定義した。また Joty ら [16] は Email などの非同期的な会話に対し HMM とグラフィカルモデルを用いた手法で教師なしの分類を、Crook ら [17] はディリクレ過程混合モデルを用いて発話のクラスタリングを行っている。これらの手法は、クラスタとしてそれぞれの発話をカテゴリ化することはできるが、分散表現のようなベクトルとしてそれぞれの素性を獲得することはできない。これらのモデルでは隠れ変数を仮定し、そのパラメータを推定するような形で教師なし学習を行っている。本論文では、隠れ変数を仮定するようなモデルではなく前後の文脈を教師データとして扱い学習するようなモデルを用いて教師なし学習を行い、分散表現を生成する。これは Bengio らの Neural Language Model [18] などで提案されている学習手法を踏襲しているものであり、近辺文脈に現れる文や単語を推定するように学習を行う形式の教師なし学習のモデルである。

2.2 文の分散表現生成

提案ニューラルネットワークでは Neural Language Model [18] や、word2vec [10] にて用いられている単語の分散表現生成手法を文レベルに拡張したモデルを用いる。単語の

分散表現生成を文レベルに拡張した研究は幾つか存在している。Le ら [19] は本提案とは別の手法を用いて word2vec を文レベルに拡張し、文の分散表現を生成する手法を提案している。また Kiros ら [20] は Sequence-to-Sequence Model [5] と word2vec のアイデアを混合した手法を用い、分散表現を生成している。これらの手法は一文についての分散表現を生成することに主眼を置いており、対話独特の特徴を捉えることは考慮していない。本論文では word2vec のアイデアを対話に適するように改良し、多くの情報を対話文から獲得するためのニューラルネットワークを提案する。

3. 提案手法

3.1 提案手法の概要

本章では、提案するニューラルネットワーク（以降、提案手法）について説明する。図 1、図 2 に提案手法の概略図を示す。図中の A_T は交互に行われる対話の中で、時刻 T に話者 A から発話された内容を表す。また対話は両話者が交互に発話することによって行われるものとし、時刻は発話が行われる度に値が +1 されるものとする。

提案手法は 2 種類のモデルから構成される。一つは図 1 に示すように、前の 2 つの発話文からその後に来る発話文を予測するように学習を行うモデルである。もう一つは図 2 に示すような特定の発話文から前後の発話文を予測するように学習を行うモデルである。上記の 2 種類のモデルに対して、対話コーパスを用いて学習を行う。学習の際には入力文に対して形態素解析を行い、単語を順に LSTM に入力する。末尾の単語まで入力したら、その際のメモリセル内 (3.2 節にて後述) のベクトルを抽出、出力する。これは図 1 および図 2 の塗りつぶされている部分にあたり、抽出の処理は白抜き矢印にあたる。また図 1 中の $\oplus$ はベクトルを連結させる処理を表す。メモリセル内のベクトルは、それぞれの Bag-of-Words を予測するように全結合のニューラルネットワークを用いて変換される。これを図 1、図 2 にある矢印の逆向きに誤差逆伝播していくことで学習が行われる。学習が終了した後は図 1、図 2 に示されている $LSTM_1$, $LSTM_2$, $LSTM_3$ を取り出し、それぞれエンコーダーとして使用する。未知の対話文に対する分散表現の生成手法を図 3 に示す。図 3 中の $\oplus$ は図 1 と同様、ベクトルを連結させる処理を表す。また図 3 中の白抜き矢印も図 1、図 2 と同様にメモリセルから抽出されたベクトルを表す。学習を行っていない話者 C と話者 D の連続する 3 発話をそれぞれ 3 種類のエンコーダーに入力し、それぞれの発話文の末尾単語が入力された際の LSTM のメモリセル内のベクトルを結合させた値を分散表現として用いる。

以下、構成要素や学習方法の詳細を説明する。

3.2 Long Short-Term Memory

LSTM は時系列処理が可能な再帰型ニューラルネットワークのモデルのひとつである。RNN における勾配消失問題 [21] を解決するために改良が施されており、近年音声認識やキャプション生成等で大きな成果を挙げている [22, 23]。

提案手法で用いた LSTM の構造を図 4 に示す。図 4 におけ

図1 提案ニューラルネットワーク（後方の発話を予測する学習）

図2 提案ニューラルネットワーク（前後の発話を予測する学習）

図3 未知の発話文に対する分散表現生成

図4 LSTMの構造

数は以下の様に定義されている.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (2)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (4)$$

$$h_t = o_t \tanh(c_t) \quad (5)$$

る x_t は時刻 t の際の入力ベクトルを表し, c_t は時刻 t におけるメモリセルの内部状態を表す. メモリセルの内部状態は時系列データの記憶を行う部分に相当している. またそれぞれの関

ここで, W はそれぞれの層間の重み行列を表し, b はそれぞれの層におけるバイアス項を表す. また σ はロジスティックシグモイド関数であり, 以下の関数を表す.

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \quad (6)$$

3.3 提案モデルの学習

LSTM の学習は Back Propagation Through Time (BPTT) [24] を用いて行う。これは時系列状態に展開した時系列データを逆方向に誤差を伝播させて学習を行う手法である。

学習における出力は Bag-of-Words と一致することを目的としている。そのため目的関数 E は以下の式で示される交差エントロピー誤差を用い、これが最小となるように学習を行う。

$$E = - \sum_k t_k \log y_k + (1 - t_k) \log(1 - y_k) \quad (7)$$

ここで y_k は出力データの k 次元目、 t_k は教師データの k 次元目を表す。

3.3.1 後方の発話文を予測する学習

後方の発話文を予測する学習は図 1 に示されているように行われる。図 1 を例とすると、1 つ前に話者 B により発話された B_{T-1} と 2 つ前に話者 A により発話された A_{T-2} の 2 つの発話文を LSTM に入力しエンコードを行う。エンコード後の分散表現から A_T の Bag-of-Words を推定するような学習が行われる。このような構造を用いて学習を行うことにより、「次に相手が発話すると推測できる発話文」が似ている発話文同士の分散表現を類似させることが出来る。また $T-2$ の発話まで考慮することで、話し相手の発話に対してのみではなく 1 つ前の自分の発話まで考慮した学習を行う事が可能になる。

3.3.2 前後の発話文を予測する学習

前後の発話文を予測する学習は図 2 に示されるように行われる。図 2 を例とすると、 A_T を LSTM に入力しエンコードを行う。エンコードされたベクトルを入力として、 B_{T-1} と B_{T+1} の Bag-of-Words を推定するように学習を行う。このようにすることで、「前後に現れる発話文」が類似している文についてはエンコード後の分散表現も類似させることが可能となる。

3.3.3 モデルの非対称性

図 1 のモデルは次の文が類似するように学習を行うのに対して、図 2 のモデルでは前後に現れる文が類似するように学習を行う。2 つのモデルで目的とする文脈が対称になっていないのは、より多くの情報を獲得するためである。これらの構造を対称とした場合、学習は同一の情報を基に行われることになる。しかし目的なる文脈を異なるものにすることで異なる特徴から学習が行われ、より複雑な文脈情報を学習できるようになると考えられる。今回の提案手法ではより多くの情報を対話文から取得するために、図 1、図 2 に示すような 2 種類の非対称なネットワーク構造を用いて文脈の推定を行っている。

4. 評価実験

4.1 評価実験の概要

提案手法により作成した分散表現に対して、2 種類の評価実験を行った。一つは対話破綻検出チャレンジ [25] のデータ

表 1 対話破綻検出で使用した実験データ

	○ (%)	△ (%)	× (%)	総発話数
学習用データ	8,520 (74.0)	1,107 (9.6)	1,879 (16.3)	11,506
開発用データ	143 (65.0)	29 (13.2)	48 (21.8)	220
評価用データ	553 (62.8)	129 (14.7)	198 (22.5)	880

を用いた実験、もう一つは対話行為タグの推定を行う実験である。

これらの実験は分散表現を素性として用いた場合の精度を比較するための実験である。精度の比較にあたり、分類器は 4 層ニューラルネットワークを使用した。多層ニューラルネットワークは任意の非線形関数を近似可能であり、一般的な分類器の一つとして使用することが可能である。予備実験より、4 層のニューラルネットワークの中間層を使い、そのノード数は 2000、学習率は 0.01 とし、勾配降下における更新規則として AdaGrad [26] を用いた。

また、LSTM への単語の入力は、単語各単語の分散表現が用いられる。それぞれの単語の分散表現は word2vec [10] を用いて生成した。その際の学習には日本語版 Wikipedia を用い、入力単語の次元数は 200 とした。また、LSTM のメモリセルや出力の次元数も 200 次元とした。学習におけるパラメータとしては、学習率 0.01、学習回数は 10,000 回、ミニバッチのサイズは 64、勾配のクリッピングは 5 とした。また勾配降下における更新規則は AdaGrad [26] を用いた。今回の実験における文の形態素解析には MeCab [27] を用いた。形態素解析のための辞書についてはより多くの固有表現への対応のため ipadic-neologd [28] を使用した。以下、それぞれの実験方法と結果、考察について述べる。

4.2 対話破綻検出

対話破綻検出はユーザーとシステムとの対話において、システムの発話が対話の流れにおいて適切であるかどうかの推定を行う。データは Docomo の雑談対話 API [30] とユーザーの発話が交互に行われており、システムの発話に以下のラベルが付与されているコーパスを用いる。

○：破綻ではない

△：破綻とはいき切れないが、違和感を感じる発話

×：あきらかにおかしいと思う発話、破綻

対話破綻検出に用いた実験用データの数を表 1 に示す。一つの発話には複数人によるアノテーションが行われているが、ここでは [25] で行われている基準と同様に、最も多く付与されているラベルを正解ラベルとして扱った。ただし、これも [25] と同様に閾値パラメータ t を 0.5 として多数ラベルの占める割合が閾値パラメータを上回る場合のみを正解ラベルとして採用し、そうでない場合は非破綻を正解ラベルとした。

4.2.1 実験条件

分散表現生成のための学習は、2016 年 3 月から 2016 年 5 月にかけて Twitter 上から取得した 20 万組の対話データを用いた。Twitter での発話データを使用した理由としては、以下の

表2 実験において使用したコーパスの一文あたりの単語数・文字数の比較

	学習用 Twitter データ	対話破綻検出用コーパス	対話行為推定用コーパス
一文あたりの単語数	12.82	7.33	13.05
一文あたりの文字数	22.97	12.61	21.93

ものが挙げられる。

1. Twitter を利用することで多量の対話データを取得することができる。
2. 今回の評価実験において使用する対話破綻検出、対話行為推定のコーパスは雑談対話であり、Twitter における対話も多くの場合雑談であるため、同一ドメインのコーパスとして利用することができる。

また、今回学習に使用した Twitter での発話データと対話破綻検出用コーパス、対話行為推定用コーパスのそれぞれにおける一文あたりの平均文字数と一文あたりの平均単語数を予備実験として調査した。結果を表2に示す。表2より、Twitter 上における発話が文字数や単語数の点においては評価データと比べて大きく変わらず、学習用コーパスとして違和感なく使用できると考えられる。

このデータは2人の話者がTwitter上で行った会話を取得したもので、一組は連続した3発話から成り立っている。収集にあたってはTwitter APIを利用した。具体的な収集方法としてはAPIを用いて任意の対話対を取得し、同一ユーザー同士で3回以上連続して発話が続いているもののみを収集する方法を用いた。3章で説明した3種類のLSTMをこのデータを用いて学習させ、それぞれのLSTMを用いて対話破綻検出用のデータに対して分散表現を生成した。比較手法として、以下の5つの手法を用いた。

1. ベースライン
2. Conditional Random Field (CRF) [25, 29]
3. 双方向 LSTM-RNN を用いた手法 [25, 29]
4. Bag-of-Words
5. Paragraph Vector [19]

ベースラインは表1の評価用データ中で最も大きな割合を占めている○ラベルの割合(62.8%)を用いた。CRFを用いた手法と双方向LSTM-RNNを用いた手法は既存研究[25, 29]中でのベースラインと最高精度の手法である。Paragraph Vectorは2章で紹介した文の分散表現を生成する為の手法の代表例として、今回比較手法の一つとした。またBag-of-Wordsは疎なベクトルで文を表現する基本的な手法である為、今回比較手法として採用した。Paragraph VectorとBag-of-Wordsはそれぞれを素性として使い、提案手法と同様の分類器で学習を行った際の精度を比較した。ここでParagraph Vectorの学習には提案手法と同様のTwitterのデータを用いた。

提案手法、Paragraph Vector、Bag-of-Wordsについては表1の学習用データを用いて4層ニューラルネットワークの学習を行い、表1の開発用データを用いて学習回数を決定した。図5に開発用データを用いた際の学習回数と精度の関係を示す。

学習回数に関しては、図5において精度が最も高い学習回数を使用した。ただしParagraph VectorとBag-of-Wordsに関し

図5 開発用データを使用した際の精度

ては過学習や未学習を避けるためにそれぞれ予備実験を行い、学習回数を決定した。Paragraph-Vectorについては、学習初期は未学習である可能性が考えられた為それぞれの学習回数と $\Delta + \times$ の際のPrecisionとRecallの値を予備実験として調査した。結果を図6に示す。図6より学習回数4回目まではほとんど $\times$ または Δ の出力がないことがわかる。それ以降で最も精度が高いのは学習回数5回目であるが、学習回数4回目まではほぼ未学習であった為学習回数5回目に十分な学習が行えているかは疑問の余地が残る。これらの理由を考慮し、Paragraph Vectorについては学習回数5回目の際の評価データによる実験結果と、学習回数6回目以降で最も精度の高い学習回数36回目の際の評価データによる実験結果を比較し、より精度の高い学習回数36回目の際の結果を実験結果として使用することとした。

またBag-of-Wordsについては、予備実験として開発用データを用いて評価を行った際の学習時の誤差の値と評価時の誤差の値を計測し、またそれらの平均値を算出した。結果を図7に示す。図7より学習回数5回目の際に平均値が最も小さな値をとっていることがわかる。以上より、Bag-of-Wordsに関しては学習回数5回目の際の評価データによる結果を実験結果として使用することとした。

4.2.2 実験結果・考察

実験結果を表3に示す。 $(\times)$ とついている項目は表1におけるラベル $\times$ についての値であり、 $(\Delta + \times)$ とついている項目は Δ と $\times$ を同一ラベルと考えた際の値を示している。精度については提案手法を用いたものが最も高くなった。またラベル $\times$ に関するPrecisionも今回比較した手法の中では提案手法が最も高い値であった。 $\Delta + \times$ のF値やPrecision、Recallに関して

表3 対話破綻検出の実験結果

	Accuracy	Precision (×)	Recall (×)	F (×)	Precision ($\Delta + \times$)	Recall ($\Delta + \times$)	F ($\Delta + \times$)
ベースライン	0.628	-	-	-	-	-	-
CRF [25, 29]	0.481	0.355	0.136	0.197	0.773	0.628	0.693
双方向 LSTM-RNN を用いた手法 [25, 29]	0.643	0.444	0.444	0.444	0.824	0.361	0.502
Bag-of-Words	0.528	0.288	0.080	0.120	0.296	0.049	0.084
Paragraph Vector	0.615	0.440	0.167	0.242	0.526	0.156	0.241
提案手法	0.652	0.466	0.379	0.418	0.675	0.343	0.454

図6 開発用データを使用した際の Paragraph Vector における Precision と Recall

図7 開発用データを用いて Bag-of-Words を評価した際の誤差と平均

は、CRF や双方向 LSTM-RNN と比べて低い値となっている。これは提案手法を用いることで、CRF や双方向 LSTM-RNN と比べてより決定的な特徴を発話文から学習できているためと考えられる。 Δ をより正確に出力する為の改善としては、ユーザーの個人性を考慮した学習を行なう必要があると考えられる。

同じ4層ニューラルネットワークを分類器として使用し比較を行った Paragraph Vector や Bag-of-Words との比較について

は、全ての項目について提案手法が上回っている。このことから、本手法を用いることで以下のことが示唆される。

- 対話破綻における検出が高精度で行えること
- 既存の素性と比較してもより高い精度で対話破綻の特徴を捉えられていること

また今回比較手法として用いた、既存手法中の最高精度のものとは1つ前までの発話を考慮して双方向型の LSTM を使用したものである。この手法と比較しても提案手法の方が高精度を記録している。具体的に、既存手法では誤った推定を行っていたが提案手法を用いることで正しい推定が行えるようになった例として以下のものが挙げられる。

対話例 1

ユーザー： 高貴な雰囲気という、具体的にはどんな人をイメージしますか？

システム： 食事が静かにできます

ユーザー： 子供のころは食事中は静かに食べなさい、としつけられたものです

システム： 食事に行きます

上記の対話において、「食事に行きます」というのは文単体では破綻しておらず、また食事の話をしているという大きな流れの中では特に問題のない発話に見える。しかし対話の前後文脈を細かく解析すると、この文脈の中では話の筋が通らない発話となっていることがわかる。このような発話に関して提案手法を用いて過去の文脈の考慮に重点を置いた学習を行うことで、より頑健な推定が行えるようになると思われる。

対話例 2

ユーザー： 今日は暑いので素麺を食べようと思っています

システム： 料理は豪華ですね

ユーザー： たしかに具がたくさんあれば豪華です

システム： 料理は家で作ってるのですか？

このような文脈における発話についても、提案手法を用いることで破綻していないという正しい推定を行うことが可能となる。直前のユーザーの「たしかに具がたくさんあれば豪華です」という発話とそれに対するシステムの「料理は家で作っているのですか？」という発話だけを見ると一見対話が成り立っていないように見えるが、その前のシステムの「料理は豪華ですね」という発話まで考慮に入れることで対話が破綻していな

表4 簡易 DAMSL タグ [31] の種類と割合

タグの種類	割合 (%)
意味を持たない発話	9.1
話者に向かう発話	0.7
第三者に向かう発話	1.1
思ったこと, 説明など	41.4
質問	26.6
命令	4.4
提案	2.6
予定	0.2
挨拶	0.4
感謝	0.5
謝罪	1.2
同意	7.1
理解	2.6
引用, 曖昧な発話	2.0

いという推定を行うことが可能となっていると考えられる。

このことより、2つ前の発話までを考慮することで、対話の文脈に関してより重要な特徴を学習できると考えられる。また提案手法では学習の構造を拡張することで3つ前、4つ前の発話を考慮することも可能である。

4.3 対話行為タグ推定

本実験ではそれぞれの発話に対して付与されている対話行為タグの種類を推定するタスクを行う。この推定の精度を測ることで、提案手法の、対話行為推定に関する素性としての表現能力を評価する。

4.3.1 実験用コーパスの作成

対話行為推定の実験を行うにあたり、日本語のタグが付与されたコーパスが存在していなかったため、作成を行った。作成にあたり、タグは磯村ら [31] により定義された簡易 DAMSL タグを使用した (表4)。これは 48 種類のタグが存在する SWBD-DAMSL タグ [12] を簡易化し、14 種類まで絞ったものである。また対話文については、インターネット上に公開されている声劇台本サイト [32] より取得した 1,127 発話を用いた。それぞれ発話文に対して簡易 DAMSL タグを、1 発話につき 1 つ付与することで、コーパスとして使用した。

作成したコーパスにおける、それぞれのタグの占める割合を表4に示す。

4.3.2 実験条件

対話破綻検出の際と同様に 20 万組の Twitter データを使用して LSTM を学習させ、使用するコーパスの発話に対して分散表現を生成した。比較手法としては Paragraph Vector と Bag-of-Words を用い、分類器は同様に 4 層ニューラルネットワークを使用した。また 14 種類のタグのうち、ランダムで 1 つのタグを選択した際の正答率をベースラインとする。このベースラインの試行は 10 回繰り返して行い、その平均値を用いた。評価については 10 分割の交差検定を行い、その平均値を用いた。

表5 対話行為推定の実験結果

手法	精度
ベースライン (ランダム)	0.071
Bag-of-Words	0.468
Paragraph Vector	0.488
提案手法	0.584

表6 対話行為推定の結果における P 値と統計検定量

比較手法	統計検定量	P 値 (両側検定)	調整後の P 値 *
Bag-of-Words	8.13	4.44×10^{-16}	8.88×10^{-16}
Paragraph Vector	5.99	2.10×10^{-9}	4.20×10^{-9}

提案手法と各比較手法との間の統計検定量を調査

* Bonferroni の方法を用いて P 値 (両側検定) を調整

図8 精度と学習回数の関係

4.3.3 実験結果・考察

提案手法、Paragraph Vector、Bag-of-Words の学習回数と推定精度を図8に示す。またそれぞれの手法の最高精度を比較したものを表5に示す。

図8より、安定して提案手法が高精度を示していることが見て取れる。また表5からはランダムも合わせた4手法の中で提案手法が最も高い精度を示していることが見て取れる。また、提案手法と比較手法に対して符号検定を行った結果と、Bonferroniの方法によって調整したP値を表6に示す。表6より、提案手法とそれぞれの比較手法との間に両側確率で有意水準1%の有意差があることが確認された。

これらの結果から、対話行為についても比較手法と比べて高い精度で推定可能であることがわかる。提案手法を用いることで推定が可能になった例として、以下のような発話について推定が可能となった。

対話例

話者1: このサラミが一杯ので?

話者2: そう、それでいいよ

話者1: 大きさはどうしますか?

話者2: M サイズ

話者 1: M でいいんですか？

話者 2: あ？

このような流れでの話者 2 の「あ？」という発話は、質問というよりは曖昧で意味を持たない発話に近いものであると考えられる。このような発話に関して提案手法を用いることで前後の文脈とその発話自体の対比をより深く見ることが可能となり、発話文の表層のみを見るような比較手法と比べて推定精度が向上したと考えられる。

上記の対話例などから、提案手法を用いることで以前の発話と今回の発話との関連性などに重点を置いて学習を行うことができるようになり、深い対話文の特徴を把握できるようになると考えられる。これらのことから、提案手法を用いることで比較手法と比べて対話文をよりの確に表現できることが示唆される。

5. おわりに

本論文では、対話における文脈を考慮した分散表現生成ニューラルネットワークを提案した。まず word2vec で用いられている 2 種類のニューラルネットワークを改良を行い、それらを組み合わせることで対話文に対して分散表現を生成できる新たなニューラルネットワークを構築した。

評価実験では対話破綻検出と対話行為推定の 2 つのタスクで行った。提案ニューラルネットワークによって作成した分散表現を素性として簡単な分類器を用いて各タスクで実験を行った。その結果として対話破綻では CRF と LSTM-RNN よりも高い精度が得られた。また 2 つのタスク両方において、同じ分類器を使用して Bag-of-Words, Paragraph Vector のそれぞれを素性として用いた場合と比較して高い精度が得られた。複数のタスクにおいて高精度を記録できたことから、本提案ニューラルネットワークによって得られた分散表現は、対話における汎用的な素性であることが示唆される。

今後の発展としては、提案モデルの学習時の予測文脈を拡張・変更させることでより多くの素性を獲得できる可能性が考えられる。

参 考 文 献

- [1] Joseph Weizenbaum. Eliza – a computer program for the study of natural language communication between man and machine. *Commun. ACM*, Vol. 9, No. 1, pp. 36–45, 1966.
- [2] Rafael E. Banchs and Haizhou Li. Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pp. 37–42. Association for Computational Linguistics, 2012.
- [3] Alan Ritter, Colin Cherry, and William B Dolan. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 583–593. Association for Computational Linguistics, 2011.
- [4] Oriol Vinyals and Quoc V. Le. A neural conversational model. *CoRR*, No. abs/1506.05869, 2015.
- [5] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014.
- [6] Jason D. Williams and Steve Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, Vol. 21, No. 2, pp. 393–422, 2007.
- [7] Yasuhiro Minami, Akira Mori, Toyomi Meguro, Ryuichiro Higashinaka, Kohji Dohsaka, and Eisaku Maeda. Dialogue control algorithm for ambient intelligence based on partially observable markov decision processes. *Proc. ISCA IWSDS*, pp. 254–263, 2009.
- [8] 目黒豊美, 東中竜一郎, 堂坂浩二, 南泰浩. 聞き役対話の分析および分析に基づいた対話制御部の構築. 情報処理学会論文誌, Vol. 53, No. 12, pp. 2787–2801, Dec 2012.
- [9] Iulian Vlad Serban, Alessandro Sordani, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Feb 12–17, pp. 3776–3784, 2016.
- [10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, No. abs/1301.3781, 2013.
- [11] J. L. Austin. *How to Do Things with Words*. Oxford, 1962.
- [12] D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical Report Draft 13, University of Colorado, Institute of Cognitive Science, 1997.
- [13] David B. Bracewell, Marc T. Tomlinson, and Hui Wang. Identification of social acts in dialogue. In *COLING*, pp. 375–390, 2012.
- [14] Elina Zarisheva and Tatjana Scheffler. Dialog act annotation for twitter conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 114–123. Prague, Czech Republic Association for Computational Linguistics, 2015.
- [15] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 172–180. Association for Computational Linguistics, 2010.
- [16] Shafiq Joty, Giuseppe Carenini, and Chin-Yew Lin. Unsupervised modeling of dialog acts in asynchronous conversations. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, Vol. 22, p. 1807, 2011.
- [17] Nigel Crook, Ramon Granell, and Stephen Pulman. Unsupervised classification of dialogue acts using a dirichlet process mixture model. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 341–348. Association for Computational Linguistics, 2009.
- [18] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, Vol. 3, No. Feb, pp. 1137–1155, 2003.
- [19] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, Vol. 14, pp. 1188–1196, 2014.
- [20] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pp. 3294–3302, 2015.
- [21] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In Kremer and Kolen, editors, *A Field Guide to Dynamical Recurrent Neural Networks*, IEEE Press, 2001.
- [22] Hasim Sak, Andrew W. Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *CoRR*, No. abs/1402.1128, 2014.
- [23] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*,

- No. abs/1411.4555, 2014.
- [24] Paul J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, Vol. 78, No. 10, pp. 1550–1560, 1990.
- [25] 東中竜一郎, 船越孝太郎, 小林優佳, 稲葉通将. 対話破綻検出チャレンジ. 言語・音声理解と対話処理研究会, Vol. 75, pp. 27–32, Oct 2015.
- [26] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, Vol. 12, No. Jul, pp. 2121–2159, 2011.
- [27] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proc. of EMNLP*, pp. 230–237, 2004.
- [28] Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab, 2015.
- [29] Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2016.
- [30] NTTdocomo 雑談対話 API. https://www.nttdocomo.co.jp/service/developer/smart_phone/analysis/chat/.
- [31] 磯村直樹, 海不二夫, 石井健一郎. HMM による非タスク指向型対話システムの評価 (自然言語処理). 電子情報通信学会論文誌. D, 情報・システム, Vol. 92, No. 4, pp. 542–551, Apr 2009.
- [32] らむりて台本部. <http://gramteril.blog.fc2.com/blog-entry-22.html>.

(2016 年 10 月 20 日 受付)

(2017 年 3 月 3 日 採録)

[問い合わせ先]

〒223-8522 横浜市港北区日吉 3-14-1

慶應義塾大学大学院理工学研究科開放環境科学専攻

三上 佳孝

TEL: 045-563-1151

E-mail: mikami@soft.ics.keio.ac.jp

著者紹介

みかみ よしたか
三上 佳孝 [非会員]

2016 年慶應義塾大学理工学部卒業. 現在, 同大学大学院修士課程在学中. 対話システムの研究に従事.

はらわら まさふみ
萩原 将文 [正会員]

1982 年慶大・工・電気卒. 1987 年同大学院博士課程修了. 工博. 同年同大助手. 現在, 同大教授. 1991-92 年度スタンフォード大学訪問研究員. 視覚・言語・感性情報処理とその融合の研究に従事. 1990 年 IEEE Consumer Electronics Society 論文賞, 1996 年日本ファジィ学会著述賞, 2004 年, 2014 年日本感性工学会論文賞, 2013 年日本神経回路学会最優秀研究賞受賞. 2015-2016 年日本知能情報ファジィ学会会長.

Distributed Representation Generation Neural Network Considering Conversational Context

by

Yoshitaka MIKAMI and Masafumi HAGIWARA

Abstract:

In natural language processing, estimation of the meaning of a sentence is one of the important issues. Various mathematical expression methods have been proposed for declarative sentences such as syntactic information and distributed representation like word2vec. However, mathematical expression methods considering dialogue context have not been proposed. This paper proposes new neural networks to estimate the features of dialogue context as distributed representation. We improve two kind of conventional neural networks employed in word2vec for sentences, and combine them to generate distributed representation. In the first neural network, Long Short-Term Memory (LSTM), one of recurrent neural networks, is employed as an encoder and learning to estimate the words (Bag-of-words) contained in the previous and the next utterances of an input utterance is carried out. In the second neural network, several LSTMs are employed as an encoder and learning to estimate the words (Bag-of-words) contained in the next utterance is carried out. Encoders in both of the neural networks are combined to construct a new neural network to generate the distributed representation for new input utterance. We carried out two kinds of experiments to evaluate proposed networks: dialogue breakdown detection and dialogue act prediction. In both of the experiments, the proposed network shows higher accuracy than the conventional methods.

Keywords: neural network, natural language understanding, distributed representation, dialogue act estimation, Long Short-Term Memory

Contact Address: **Yoshitaka MIKAMI**

Open and Environmental Systems, Graduate School of Science and Technology, Keio University

3-14-1, Hiyoshi, Kohoku-ku, Yokohama City 223-8522, Japan

TEL: +81-45-563-1151

E-mail: mikami@soft.ics.keio.ac.jp